Modeling Filled Pauses and Silences for Responses of a Spoken Dialogue System

Kengo Ohta, Norihide Kitaoka, Seiichi Nakagawa

***Abstract*—**In human-to-human dialogue, pauses such as filled pauses and silences play an important role not only as markers of discourse structures[1] but also as cues to subsequent phrases[2]. As shown in these previous literatures, the modeling of filled pause[20][21] is essential not only in implementation of speech recognition system for spontaneous speech[22][23] but also in implementation of natural spoken dialogue system, considering the effects of these phenomena upon users. In this paper, we propose the modeling of filled pauses and silences in response utterances of spoken dialogue systems. At first, the positions of pauses are investigated in corpus study in dialogue data and presentation data of the Corpus of Spontaneous Japanese (CSJ). Based on this investigation, pauses are modeled and inserted into response utterances. Our proposed method is evaluated in subjective experiments of a tourist-guiding task. We compared user comprehension, naturalness and listenability of the system's responses with and without filled pauses and silences. Our results showed that the filled pause positioned at the inter-sentence level can enhance the user comprehension and improve the naturalness of a spoken dialogue system.

***Keywords*—** Filled pause, Naturalness, Silence, Spoken dialogue system, Understandability, Listenability.

I. INTRODUCTION

Filled pauses and silences in human speech play an important role in human dialogue. Donzel and Beinum[3] investigated patterns of filled pauses and silences in Dutch speech. Their results showed that pausing in discourse is achieved by silent pauses, filled pauses, lengthening of words, and combinations of these three phenomena. Swerts et al.[1] analyzed filled pauses in a spontaneous monologue and showed that these pauses carried information about discourse structures. Moniz et al.[4] conducted a corpus study on prepared (non-scripted) and spontaneous oral presentations in a high school context. They showed that filled pauses and segmental prolongations performed similar functions and may be considered to be in complementary distribution, while obeying general syntactic and prosodic constraints. Watanabe et al.[2] examined the occurrence of filled pauses and silences as cues to subsequent phrases. Their research showed that the duration of filled pauses and silences gave listeners cues to the complexity of upcoming phrases. Somiya et al.[5] reported how filled pauses used in lectures influenced the understanding and listening ability of the audiences. Based on this investigation, Naito et al.[6] proposed a decision-tree-based advice system for improving the usage of filled pauses in a lecture context. As shown in these studies, filled pauses and silences affect the listenability of speech and listeners' comprehension. These effects should be considered in the implementation of natural spoken dialogue systems. For instance, Itoh et al.[7] analyzed the use of filled pauses in a dialogue system. Their investigation supports the effectiveness of inserting filled pauses between sentences as a sign of the ongoing activity of a spoken dialogue system when it requires a long time to produce the next sentence. Similarly, Shiwa et al.[8] reached the same conclusion in human-robot interactions. The "two second rule" is a well-known finding in previous studies of user interfaces: a system should not take more than two seconds after input to respond[9][10]. All these results indicate that it is important to consider the usage of filled pauses and silence when designing a spoken dialogue system. Additionally, Itoh et al.[11] revealed that the frequency of filled pauses used by a human changes depending on whether he/she is talking with machine or another human. Adel et al.[12] proposed a speech synthesis model based on the definition of disfluency and the concept of underlying fluent sentences. Evaluations of their system showed that it was impossible to synthesize filled pauses without decreasing the overall naturalness of the system.

In this study, we model filled pauses and silences in the responses of spoken dialogue systems. Particularly, we focus on the pauses at the inter-sentence level, which have not been studied in most of the previous studies. In subjective experiments of a tourist-guiding task, we compared user comprehension, naturalness, and listenability of the system's responses with and without filled pauses and silences.

The remainder of this paper is organized as follows: Section 2 provides the definitions of the filled pauses and silences we used in our experiments. Section 3 outlines the speech dialogue task used in our experiments. Section 4 describes settings and results of our subjective experiments. Section 5 outlines our conclusions.

II. POSITION AND LENGTH OF PAUSE

Silences and filled pauses are produced by two main factors.

Kengo Ohta is with the Dept. of Systems and Control Engineering, Anan National College of Technology, 265 Aoki Minobayashi, Anan, Japan (e-mail: kengo@anan-nct.ac.jp).

Norihide Kitaoka is with the Grad. School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan (e-mail: kitaoka@nagoya-u.jp).

Seiichi Nakagawa is with the Dept. of Computer Sciences and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka Tenpaku-cho, Toyohashi, Japan. (e-mail: nakagawa@slp.cs.tut.ac.jp).

The first is physiological factors such as breathing. Pauses caused by these factors are produced at positions that are irrelevant to the linguistic breaks. The second factor is the planning strategy of utterances, where pauses are produced at the boundaries of information units. Minematsu and Nakagawa[13] carried out an analysis of the correlation between acoustic and perceptual pauses (human perceived linguistic breaks). Their investigation of interjections and filled pauses around the perceptual pauses showed that there was a high probability that the acoustic properties of えー(”ee”), えーと(”eeto”) and で(”de”) caused the perceptual pauses. Kitahara et al[18] investigated the effects of prosody and silence on listener's comprehension. Their results revealed that prosody make listeners easier to understand the contents of speech while at some concentrative work. They concluded that silences take an important role as the durational buffer for listener's cognition especially in distracted condition. Watanabe[19] made two hypotheses about filled pause: (1) filled pause tends to occur at the linguistic break such as phrase boundary, clause boundary, sentence boundary and discourse boundary. The frequency of filled pause increases according to the strength of break (boundary hypothesis). (2) The frequency of filled pause increases when the succeeding phrase or clause has complex structure (complexity hypothesis). These hypotheses are supported by the corpus study of the academic presentation and the simulated presentation in the corpus of spontaneous Japanese (CSJ). According to Watanabe's study, at the strong break, filled pauses occur at the same rate as at the sentence boundary. Watanabe also investigated the effects of filled pauses and silences on listener's comprehension. In Watanabe's experiment, subject listens the explanation about the shape of physical body and chooses the correct body which accord with the explanation he or she heard. As the result of this experiment, the subject's accuracy was 99.2% when a filled pause is inserted at the beginning of sentence, 97.0% when a silence is inserted at the beginning of sentence, 97.6% when no filled pause and silence is inserted. Additionally, the reaction rate of subject was the same between when a filled pause or silence is inserted. On the other hand, the reaction rate degrades when no pause is inserted. Watanabe concluded that filled pauses prompts the user's preparation of comprehension.

A. Corpus Study

We conducted the corpus study of CSJ to investigate the insertion position of filled pause. In the dialogue data of CSJ, both filled pauses and *bunsetsu* segments (a Japanese phrasal unit, which consists of at least one content word and zero or more functional words) are labeled. In this study, we analyzed the relation between filled pause and bunsetsu boundary in the speech dialogue data of CSJ. Although the clause segments are not labeled in the dialogue data, the core data of academic presentation and simulated presentation include labels of both the clause boundary and the filled pause. We also analyzed the relation between filled pause and clause boundary.

The bunsetsu segment defined in CSJ is not only the basic unit in dependency parsing, takes an important role as the linguistic and semantic segment in many applications. In our analysis, we investigated the occurrence rate of filled pause at the bunsetsu boundaries in the 58 dialogue data of CSJ. As the result of this analysis, 87.1% of filled pauses occurred at the bunsetsu boundary. Except for the filled pauses which is caused by physiological factors such as breathing, the most of filled pauses occur at the semantic boundaries.

In CSJ, clause segment is also defined as the linguistic and semantic unit. We also analyzed the occurrence rate of filled pause at the clause boundaries in the core data of academic presentation and simulated presentation. Here, the clause boundaries are divided into three categories (weak boundary, strong boundary, and absolute boundary) according to the strength of break. The result of analysis is shown in TABLE I. As shown in this table, 41.1% of filled pauses occurred at the clause boundary. We can conclude that the filled pauses tend to occur at the boundaries of semantic units.

TABLE I
RELATION BETWEEN FILLED PAUSE AND CLAUSE BOUNDARY

Clause Boundary	Occurrence Rate of Filled Pause
Weak Boundary	21.7%
Strong Boundary	9.4%
Absolute Boundary	10.0%
Total	41.1%

B. Modeling of Pause based on Corpus Study

We focus on the linguistic breaks of filled pauses and silences at the inter-sentence level. Here, we take an example of a sentence providing instructions for a tourist-guiding task.

札幌から美瑛へ行くには / 札幌駅からスーパーカムイに乗って旭川駅まで 1 時間 30 分 / 旭川駅から直行バスで美瑛駅まで 50 分です.
(To get from Sapporo to Biei / travel an hour and a half by the limited express train Super-Kamui from Sapporo station to Asahikawa station / then 50 minutes by non-stop bus from Asahikawa station to Biei station.)

In this case, four distinct components － traveling time, transportation, departure time, and destination make up a unit of information. Linguistic breaks would be inserted between such components, that is, the positions marked by ”/” in the example.

In this study, we insert filled pauses or silences at the linguistic breaks defined as the boundaries of information units described above. In their study of a dialogue system, Itoh et al.[7] compared 13 kinds of filled pauses that occurred between sentences, such as: あ(”a”), あの(”ano”), え (”e”), えー(”ee”), えと(”eto”), えーと(”eeto”), えっ と(”eqto”), じゃ(”ja”), その(”sono”), で(”de”), ま (”ma”), ま

TABLE II
TOURS ADOPTED FOR OUR EXPERIMENTS

(a) Each Scenario

No.	Part	Departure	Destination	# of Information Slots	# of Morae	# of Words	User Comprehension (%, Average)	Accuracy of 2nd Task (%)
1	1st Half	Sapporo	Yuubari	5	78	32	46.3	85.9
2			Noboribetsu	5	96	40	46.7	83.5
3			Otaru	5	92	39	32.2	88.2
4			Abashiri	8	120	47	33.9	85.9
5			Kushiro	6	120	53	29.5	71.7
6			Shiretoko	8	128	53	30.7	77.7
7			Souya	11	168	65	30.7	74.7
8	2nd Half		Asahiyama	5	96	39	58.3	82.3
9			Biei	5	83	36	52.9	77.3
10			Touyako	5	93	39	42.5	82.7
11			Hakodate	8	131	54	42.9	77.9
12			Erimo	8	133	57	28.1	83.0
13			Goryoukaku	11	164	69	31.4	80.2
14			Matsumae	14	193	78	32.4	78.8

(b) Average of 1st Half and 2nd Half

Part	# of Information Slots	# of Morae	# of Words	User Comprehension (%, Average)	Accuracy of 2nd Task (%)
1st Half	6.86	114.6	47.0	35.7	81.1
2nd Half	8.00	127.6	53.1	41.2	80.3

あ(” maa”), や(” ya”). We insert the same categories of filled pauses.

III. TASK DESCRIPTION

Todo et al.[14] are developing a spoken dialogue system with multiple agents for spontaneous conversation at two major Japanese leisure venues; Hokkaido (a snowy region) and Okinawa (a tropical region). In their dialogue, one agent recommends good points while a second agent states bad points about Hokkaido and Okinawa. It is possible to draw users into the subjective nature of the dialogue by ensuring that the agents have conflicting opinions. Moreover, strategies for arranging the different agents' opinions with the aim of drawing the user into a specific opinion can be incorporated into the system.

The conversations/speech used for our experiments are supposed to be incorporated into the dialogue system devised by Todo et al. We used a touristguiding task of Hokkaido, Japan's second largest island for our subjective experiments. The pathways of a tour from Sapporo, the center of Hokkaido, to 14 major tourist sites (as shown in Table II) are guided by an agent. These pathways were recorded using the same female speaker's voice as used in the Todo et al.[14] system.

IV. EXPERIMENTS

A. Methodology

1) Experimental Procedure

In our experiments, subjects were directed to conduct the following conversation with a 3D agent displayed on a screen. The 3D agent is created using TVML Player II Version.2.3[1] provided by NHK (Nippon Hoso Kyokai, English name: Japan Broadcasting Corporation) Science and Technical Research Laboratories. The 3D agent speaks with lip syncing and bodily expression.

Agent: I'm a tourist guiding system. I'll give a grand tour to Hokkaido.
Subject: How can I get to XXX from Sapporo?
Agent: To get from Sapporo to XXX, ... (announcement)

In our experiment, we adopted a Wizard of Oz style of interaction, where the subject thought they were interacting with an automated computer dialogue system, when in actual fact, we were manually selecting the utterances that the agent presented to the subject. At the beginning of conversation, the subject is

[1] 1http://www.nhk.or.jp/strl/tvml/english/player2/

given only the departure and destination locations (XXX at the above example). After the dialogue agent explained how to travel from the departure point to the destination, the subject was directed to draw the travel pathway, as shown in Figure 2, with as much detail as possible. Additionally, as shown in Figure 1, the subject was instructed to solve simple calculation problems during the agent 's explanation. These calculation problems were given every three seconds and involved simply adding or subtracting two single figure digits (e.g. 3 + 1, 7 − 2, 4 − 5, etc). The introduction of this second task is to add a concurrent cognitive load to the subject, making it more difficult to understand the agent 's directions[15]. Note taking or provision of answers during the agent 's explanation was not allowed.

Fig. 1 An example of agent presented in our experiments.

Fig. 2 An example of transfer pathway agent explained.

TABLE III
NUMBER OF SUBJECTS FOR EACH CONDITION

Part Condition	Filled Pause	Silence	Without Pause
First Half (1-7)	8	8	8
Second Half (8-14)	8	8	8

The subject was directed to undergo 14 sets of dialogue. The condition of utterances, that is, (i) with filled pauses, (ii) with silences, (iii) without pause, was changed between the first half and the second half of these sets. Therefore, one of the six combinations of conditions shown in Table III is given for each subject. We evaluated user comprehension, listenability, and naturalness in this experiment. The user comprehension was evaluated as an accuracy rate of the subjects 'answers regarding the travel times, transportation methods, and stopovers involved in the travel path. Listenability and naturalness were evaluated on a five point scale. For example, naturalness in each half of the dialogue sets was evaluated according to the following scale:

- 5: utterances in the first half were quite natural.
- 4: utterances in the first half were relatively natural.
- 3: There was no difference between utterances in the first half and the second half.
- 2: utterances in the second half were relatively natural.
- 1: utterances in the second half were quite natural.

The two halves of the dialogue sets were then analyzed via pairwise comparison. The 24 subjects were students of the Anan National College of Technology. The number of subjects for each condition is shown in Table III.

2) Preparation of Agent's Utterances

The agent's utterances are speech recordings by a female member of a drama group at Nagoya University. Two kinds of utterances, that is, "without pause" and "with filled pause" are independently recorded for No.1～14. The "without pause" utterances are recordings of the read speech of announcement text without any filled pause or silence. However, it is noted that some short silences which occurs in fluent read speech are actually included in these utterances. On the other hand, the "with filled pause" utterances are recordings of the read speech of announcement text with filled pauses inserted. Examples of announcement text for "without pause" utterance and "with filled pause" utterance are shown below.

- **Without pause:** To get from Sapporo to Biei / travel an hour and a half by the limited express train Super-Kamui / from Sapporo station to Asahikawa station / then 50 minutes by non-stop bus / from Asahikawa station to Biei station. (*Actually, short silences are inserted at the positions marked by "/".*)
- **With filled pause:** To get from Sapporo to Biei ***uh*** travel an hour and a half by the limited express train Super-Kamui from Sapporo station to Asahikawa station ***um*** then 50 minutes by non-stop bus from Asahikawa station to Biei station.

The "with silences" utterances are generated by replacing filled pauses in "with filled pause" utterances with silences. The rates of utterances were a uniform 8 mora/sec, defined by a Time Domain Harmonic Scaling (TDHS) based time scale modification tool PICOLA [2]. The mora is a Japanese sub-syllabic unit which can be a vocalic nucleus (V) or a nucleus plus syllabic onset (CV). A CRF-based Japanese morphological analyzer MeCab ver. 0.963[3] (with Uni-Dic ver. 1.3.12[4]) was used to calculate the number of words.

The list of travel pathways used in the experiment and the statistics of announcement utterances are shown in Table II. Here, "# of Information Slots" refers to the number of information slots such as travel time, transportation method, and stopovers which a subject receives in the experiments. "Acc. of 2nd Task" indicates the accuracy rate of the calculation problems (second task). As shown in this column, the difficulty

[2] http://keizai.yokkaichi-u.ac.jp/~ikeda/research/picola.html
[3] http://mecab.sourceforge.net/
[4] http://www.tokuteicorpus.jp/dist/

of second task varies slightly among No.1～14. Although the 2nd task was easy, the accuracy rate was not so high (about 80%). This indicates that the 2nd task gave the cognitive loard for understanding the anouncement as expected. Owing to the effect of this cognitive load, the average rate of the user comprehension (1st task) was low (about 40%). Additionally, as shown in Table II (b), there was little practical difference between the difficulty of the 1st half and the 2nd half.

As stated above, some short silences which are produced in read speech occurred in "without pause" utterances. The lengths of silences in each utterance are shown in TABLE IV. The distribution of length of silences in "without pause" utterances are also shown in Fig.3. As shown in the Fig.3, some silences at the length of from 0.3 sec to 0.5 sec occurs in each utterance.

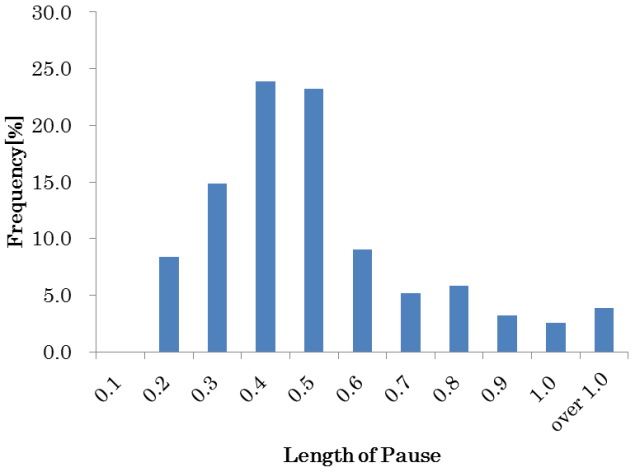

Fig. 3 Distribution of Pause Length in "without Pause" Utterances

TABLE IV
PAUSE LENGTH IN "WITHOUT PAUSE" CONDITION

No.	Destination	Whole Durations [s]	Duration of Pauses	Pause Ratio(%)
1	Yuubari	12.7	2.9	22.8
2	Noboribetsu	15.7	4.1	26.4
3	Otaru	16.8	4.4	25.9
4	Abashiri	23.2	4.6	20.0
5	Kushiro	23.0	4.8	21.0
6	Shiretoko	24.0	6.5	27.1
7	Souya	30.6	6.5	21.1
8	Asahiyama	16.7	2.9	17.5
9	Biei	14.3	3.1	21.4
10	Touyako	16.1	3.6	22.2
11	Hakodate	23.0	4.6	19.9
12	Erimo	23.2	6.3	27.2
13	Goryoukaku	31.9	8.0	25.0
14	Matsumae	37.0	9.1	24.7

B. Experimental Results

1) Effects of pauses

The experimental results of user comprehension, listenability, and naturalness are shown in Table V, Figure 4, and Figure 5.

As shown in Table V, the highest user comprehension was achieved at the "With Filled Pause" condition. This suggests that the filled pause helps users to comprehend the content of sentences. On the other hand, contrary to our intuition, silence slightly impeded users' comprehension. In a post-experimental survey, some subjects responded that it was difficult to perceive the end of sentences when silences were inserted. This indicates that silences can confuse users and impede their understanding and integration of information.

TABLE V
EFFECTS OF EXISTENCE OF PAUSE

Condition	Without Pause	With Filled Pause	With Silence
User Comprehension(%)	36.8	43.4	34.0

Some mixed results with regard to listenability are shown in the Figure 4. The subjects' answers are clearly divided with respect to the comparison between "Without Pause vs With Filled Pause", which is contrary to the results of user comprehension. This suggests that users' performance does not necessarily correspond to their intuition. On the other hand, the comparison of "With Filled Pause vs With Silence" showed a similar pattern to the results of user comprehension.

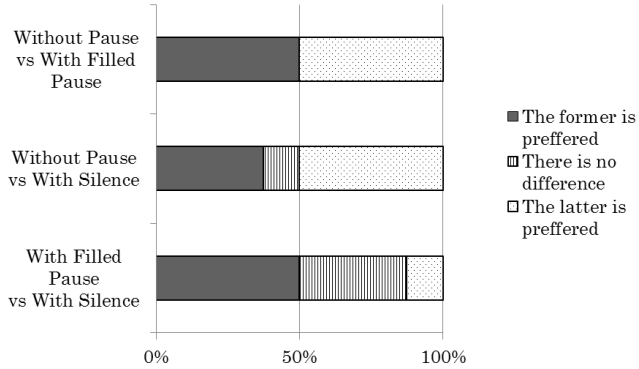

Fig. 4 Effects of Existence of Pause on Listenability.

The results of naturalness as shown in Figure 5 corresponded to our intuitive conclusions. That is, the highest naturalness was achieved when filled pauses were inserted. Additionally, there is little difference between the "Without Pause" and "With Silence" conditions. However, it is noted that the naturalness of sentences with filled pauses is strongly affected by the naturalness of the voicing of the filled pause itself. In a preliminary experiment using recordings of utterances from a person with no dramatic or elocution experience, the naturalness of filled pauses inserted into utterances was very low. As Adel et al.[12] suggests, it is difficult to produce a "natural" filled pause with a modern speech synthesizer.

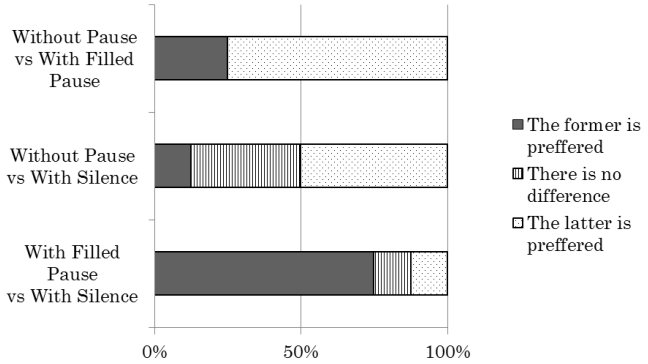

Fig. 5 Effects of Existence of Pause on Naturalness.

2) Effects of length of pause

We investigated the relationship between the duration of pauses and user comprehension. The whole duration and pause (filled pause and silence) ratio of each utterance is shown in Table VI. The correlation between the ratio of filled pauses and user comprehension was -0.01, and that between the ratio of silence and user comprehension was -0.10. As shown in these results, no correlations are observed for both types of pause. The average ratio of No.1～No.7 and No.8～No.14 was 5.1% and 5.2% respectively. There was no difference each other. We can conclude that there are no significant correlation between the duration of pauses and user comprehension.

3) Effects of frequency of pause

We also investigated the relationship between the frequency of pauses and user comprehension. As a result, the correlation between the frequency of filled pauses and user comprehension was -0.335, and that between the frequency of silence and user comprehension was -0.341. Additionally, when the frequency of pauses is normalized by the number of words in the utterance, the correlation between the frequency of filled pauses and user comprehension was -0.191, and that between the frequency of silence and user comprehension was -0.241. As shown in these results, slight negative correlations are observed for both types of pause. The frequency of pause is also shown in TABLE VI.

4) Relation between sentence length and user comprehension

We also investigated the correlation between the number of information slots, words or morae in sentences and user comprehension. The correlation between the number of information slots in a sentence and user comprehension was -0.264. Similarly, the correlation between the number of words in sentence and user comprehension was -0.288. Additionally, the correlation between the number of moras in sentence and user comprehension was -0.300. As shown in these strongly negative correlations, the longer or more complex the sentence is, the more difficult it is for users to comprehend the contents of the sentence. Also, it is logical that the more information slots which the user have to memorize, the less he/she will memorize on average (as we do not allow for taking notes). The scatter plots between user comprehension and the number of information slots, words, and morae are shown in Fig. 5, Fig. 6, and Fig. 7.

TABLE VI
DURATION OF EACH UTTERANCE

No.	Destination	Whole Durations	Duration of Pauses	Pause Ratio(%)	Pause Frequency
1	Yuubari	12.7	0.53	4.2	1
2	Noboribetsu	15.7	0.31	2.0	1
3	Otaru	16.8	0.95	5.7	2
4	Abashiri	23.2	1.77	7.6	3
5	Kushiro	23.0	1.29	5.6	2
6	Shiretoko	24.0	0.94	3.9	2
7	Souya	30.6	2.04	6.7	4
8	Asahiyama	16.7	1.34	8.0	1
9	Biei	14.3	0.83	5.8	2
10	Touyako	16.1	0.44	2.7	1
11	Hakodate	23.0	1.00	4.4	2
12	Erimo	23.2	1.01	4.4	2
13	Goryoukaku	31.9	2.30	7.2	4
14	Matsumae	37.0	1.48	4.0	3

V. Conclusion

In this paper, we investigated the effects of filled pauses and silences in a spoken dialogue system. We compared the user comprehension, naturalness and listenability of the system's responses with and without filled pauses and silences during subjective experiments of a tourist-guiding task. Our results showed that the filled pause positioned at the inter-sentence level can enhance the user comprehension and improve the naturalness of a spoken dialogue system. However, it is noted that the efficacy of inserting filled pauses depends upon the quality and naturalness of the speech synthesizer or speaker.

In future work, we are planning to explore the optimum position of filled pauses for enhancing the users' comprehension and improving the naturalness of a spoken dialogue system. Our next goal is to construct an automatic decision model for the optimum positions of filled pauses. To achieve this, we will expand upon our previously proposed

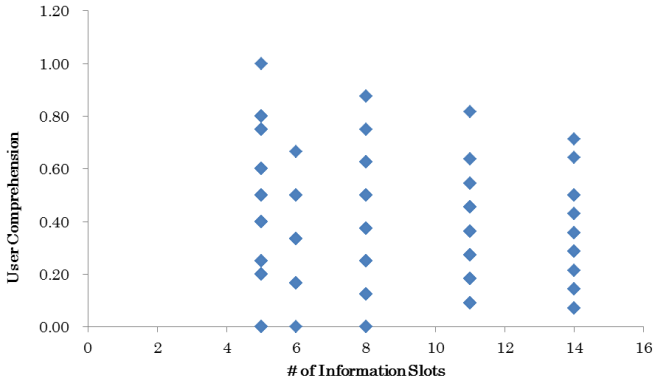

Fig. 6 Correlation between User Comprehension and Number of Information Slots

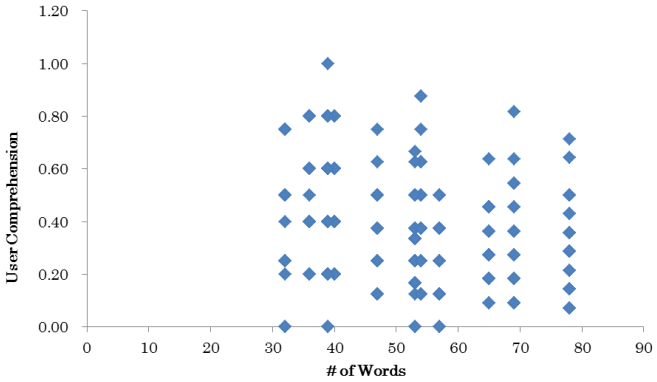

Fig. 7 Correlation between User Comprehension and Number of Words

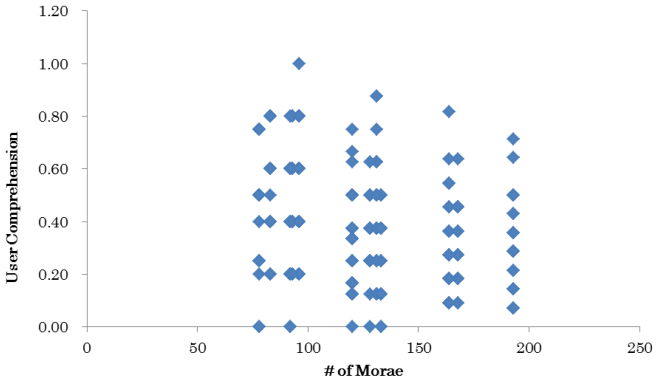

Fig. 8 Correlation between User Comprehension and Number of Morae

model for the insertion of filled pauses[16] and silence[17] for spoken dialogue systems considering longer range features. Parse trees and dependency trees will be introduced to the model to capture the important information units.

REFERENCES

[1] Marc Swerts, Anne Wichmann, and Robbrt-Jan Beun, "Filled pauses as markers of discourse structure," in Proc. of ICSLP '96, 1996, vol. 2, pp. 1033-1036.

[2] Michiko Watanabe, Keiichi Hirose, Yasuharu Den, and Nobuaki Minematsu, "Filled pauses as cues to the complexity of following phrases," in Proc. of INTERSPEECH 2005, 2005, pp. 37-40.

[3] Monique E. van Donzel and Florien J. Koopmans van Beinum, "Pausing strategies in discourse in dutch," in Proc. of ICSLP '96, 1996, vol. 2, pp. 1029-1032.

[4] Helena Moniz, Ana Isabel Mata, and M.Ceu Viana, "On filled-pauses and prolongations in eurpean portuguese," in Proc. of INTERSPEECH2007, 2007, pp. 2645-2648.

[5] M.Somiya, K.Kobayashi, H.Nishizaki, and Y.Sekiguchi, "The effect of filled pauses in a lecture speech on impressive evaluation of listeners," in Proc. of INTERSPEECH2007, 2007, pp. 2673-2676.

[6] Wataru Naito, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, "Evaluation and advice system for improving the manner of speaking in lectures using features of filled pauses," in Proc. of APSIPA ASC 2011, 2011, p. 4 pages.

[7] Toshihiko Itoh, Nobuaki Minematsu, and Seiichi Nakagawa, "Analysis of filled pauses and their use in a dialogue system," Journal of the Acoustical Society of Japan (in Japanese), vol. 55, no. 5, pp. 333-342, 1999.

[8] Toshiyuki Shiwa, Takayuki Kanda, Michita Imaia, Hiroshi Ishiguro, and Norihiro Higata, "How quickly should a communication robot respond?," International Journal of Social Robotics, vol. 1, no. 2, pp. 141-155, 2009.

[9] R. B. Miller, "Response time in man-computer conversational transactions," in Proc. of Spring Joint Computer Conference, 1968, pp. 267-277.

[10] T. Starner, "The challenges of wearable computing: Part 2," IEEE Micro, vol. 21, no. 4, pp. 54-67, 2001.

[11] Toshihiko Itoh, Atuhiko Kai, Yoshiyuki Iwamoto, Makoto Mizutani, Hiroki Yuasa, Tatsuhiro Konishi, and Yukihiro Itoh, "Comparison of linguistic and acoustic features caused by different dialogue situations in a landmark-input task," Information Processing Society of Japan Journal (in Japanese), vol. 43, no. 7, pp. 2118-2129, 2002.

[12] Jordi Adell, David Escudero, and Antonio Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," Speech Communication, vol. 54, no. 3, pp. 459 - 476, 2012.

[13] Nobuaki Minematsu and Seiichi Nakagawa, "Correlation between acoustic pauses and perceptual pauses in speech," in Proc. of ASR and ASJ Third Joint Meeting, 1996, pp. 1193-1198.

[14] Yuki Todo, Ryota Nishimura, Kazumasa Yamamoto, and Seiichi Nakagawa, "Development and evaluation of spoken dialog systems with one or two agents through two domains," in Proc. of TSD 2013, 2013.

[15] David L. Strayer and William A. Johnston, "Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone," Psychological Science, vol. 12, no. 6, pp. 462-466, 2001.

[16] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa, "Evaluating spoken language model based on filler prediction model in speech recognition," in Proc. of Interspeech 2008, 2008, pp. 1558-1561.

[17] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa, "Effective use of pause information in language modeling for speech recognition," in Proc. of Interspeech 2009, 2009, pp. 2691-2694.

[18] Yoshinori Kitahara, Syoichi Takeda, Akira Ichikawa, "Role of prosody in cognitive process of spoken language.", Journal of The Institute of Electronics, Information and Communication Engineering (in Japanese), Vol. 70, No. 11 pp.2095-2101, 1987.

[19] Michiko Watanabe. Features and Roles of Filled Pauses in Speech Communication. Hituji Syobo, 2009.

[20] A. Zgank, T. Rotovnik, and M. S. Maucec, "Slovenian spontaneous speech recognition and acoustic modeling of filled pauses and onomatopoeas", WSEAS Transaction on Signal Processing, vol. 4, no. 7, pp. 388-397, 2008.

[21] A. Zgank, T. Rotovnik, and M. S. Maucec, "Modeling Filled Pauses for Spontaneous Speech Recognition Applications", WSEAS International Conference on Application of Electrical Engineering, pp.42-47, 2008.

[22] Y. Fujii, K. Yamamoto, S. Nakagawa, "Large vocabulary speech recognition system:SPOJUS++", WSEAS International Conference MULTIMEDIA SYSTEMS & SIGNAL PROCESSING, pp.110-118, 2011.

[23] R. Thangarajan, A.M. Natarajan, M. Selvam "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language", WSEAS TRANSACTIONS on SIGNAL PROCESSING, Issue 3, Volume 4, pp. 76-85, 2008.

Kengo Ohta received his B. S. and M. S. and Ph. D degrees from Toyohashi University of Technology. He is now an Assistant Professor in Anan National College of Technology.

Norihide Kitaoka received his B. S. and M. S. degrees from Kyoto University. In 1994, he joined DENSO CORPORATION. In 2000, he received his Ph. D degree from Toyohashi University of Technology (TUT).He joined TUT as a Research Associate in 2001 and was a Lecturer from 2003 to 2006.Since 2006 he has been an associate professor in Nagoya University.

Seiichi Nakagawa received Dr. of Eng. degree from Kyoto University in 1977.He joined the faculty of Kyoto University in 1976 as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he was an Assistant Professor, from 1983 to 1990, he was an Associate Professor, and since 1990, he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Sciences, Carnegie-Mellon University, Pittsburgh, USA. He received the

1997/2001/2013 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electro, Telecomm. Engrs. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface and artificial intelligence. He is a Fellow of IPSJ and IEICE.